UNITED STATES AIR FORCE

School of AVIATION MEDICINE

DISCRIMINATORY ANALYSIS

VI. On the Simultaneous Classification of Several Individuals

PROJECT NUMBER 21-49-004
REPORT NUMBER 6

PROJECT REPORT

DISCRIMINATORY ANALYSIS

VI. On the Simultaneous Classification of Several Individuals

E.L. LEHMANN, Ph.D.
University of California, Berkeley

PROJECT NUMBER 21-49-004
REPORT NUMBER 6

USAF SCHOOL OF AVIATION MEDICINE
RANDOLPH FIELD, TEXAS

AUGUST 1951

ON THE SIMULTANEOUS CLASSIFICATION OF SEVERAL INDIVIDUALS

1. Introduction.

It has been usual in the study of classification problems to consider the classification of one item at a time. However, in practice one frequently deals with a whole group of items each of which has to be assigned to its proper category. There seem to be two main reasons why it is worth considering the problem in this more general form.

First, one may gain in efficiency. This happens, roughly speaking, because one can utilize the totality of observations to obtain estimates of unknown parameters. For certain problems this has always been realized and procedures obtained in this manner have been considered in the literature. For certain other types it was pointed out first by Robbins [1], and another example was given by Levene [2].

The other important reason for considering several items simultaneously is that one is thus led to new formulations. In particular, problems arise where the definition of the various categories is given not absolutely but in terms of the other items in the group. A suggestive although not

This work was done at Columbia University and supported in part by the USAF School of Aviation Medicine.

too typical example is the assigning of grades in an examination or course.

Throughout the present discussion we shall assume that the items to be classified are all of one kind, and that all of them are to be distributed among the same categories. Under these circumstances it is frequently reasonable to measure the loss resulting from wrong decisions by the number of incorrect classifications. (By "risk" we shall then mean the expected number of wrong classifications.) This is of course nearly always a considerable oversimplification. However, it is necessary to work with standardized loss functions, and the one suggested does have a concrete and intuitively appealing significance. It is comparable with the "simple" loss functions utilized by Wald and with the formulation of hypothesis testing in terms of the probabilities of errors given by Neyman and Pearson.

For our purpose it is important to distinguish classification problems according to the nature of the items to be classified. These may be on the one hand students, prospective doctors, skulls, plants, etc. On the other hand they may be varieties of wheat, different production processes or treatments of a disease. The division of course is not completely clear-cut. However, in the cases listed second one classifies on the basis of independent, identically distributed variables; that is, one is dealing with (statistical) populations and makes the classification on the basis of a random sample from the population.

This is usually not the case in the group of problems listed first. There the basis of classification is a set of measurements, one each of a number of different characteristics. The usual assumption attributes to this set of measurements a multivariate normal distribution. This assumption implies that the item being measured has itself been obtained by means of a chance mechanism from some population of such items; the replication of the experiment consists in drawing another batch of items from this population. Suppose now that it is desired to classify the items each into one of a number of categories. Then each item of the total population falls into one of these categories; the number in the categories is in certain proportions. It follows that for each of the items to be classified there is a definite probability of falling into each of the categories. While thus the assumption of a priori probabilities for the various categories seems inevitable when one is classifying individuals, this assumption is usually inappropriate for the classification of populations.

The description of the items in the one case as individuals, in the other as populations is of course a somewhat loose one. Thus, for example, each student in a class plays the role of a population in the following problem. The knowledge of the student is to be tested by a true-false examination. One possible measure of his knowledge is the proportion, among the totality of true-false questions that could be asked, which he can answer. The questions that are asked in the examination can be thought of (in a very rough approximation) as

a random sample from the totality of possible questions. Another example of this kind is provided when we attempt to determine the genotype of an individual through the number of recessives among his offspring.

However, generally speaking, there is a fairly clear distinction between the two types of problems. When we classify populations we are dealing with samples from distributions with unknown but fixed parameters. In the case of individuals, we assume multivariate distributions involving parameters which themselves constitute random variables.

There is a further difference between the two types of problems, which is not of a theoretical nature but which nevertheless is of some importance. Roughly speaking, and admitting that there are important exceptions, we can say that the simultaneous classification of populations usually involves only a small number, say 2 to 10, populations, while the number of individuals in a group to be classified frequently is considerably higher. Thus it is of interest to develop an asymptotic theory for the classification of a large number of individuals. On the other hand, in the classification of populations it is usually not too reasonable to assume a large number of them. An asymptotic theory here would more likely be concerned with large samples from each of the populations.

Classification problems differ not only through the nature of the objects that are to be classified but also in various respects according to the categories among which the items are to be distributed. As a first distinction, we may

either be dealing with k clearly distinct categories or with classification according to a parameter with a continuous range of variation which is broken up (sometimes somewhat arbitrarily) to provide the categories. We shall here concern ourselves mainly with the first of these two types of problems. Although this has a somewhat restricted range of applications it does arise naturally in problems of taxonomy. In particular one would expect it to be of increasing importance for the determination of genotypes, which is of interest in genetical and anthropological research. (See for example [3]).

For the problem of k categories corresponding to k distinct values of a characteristic parameter, there is a further important subdivision according to the amount of information that is available concerning these categories. In the simplest case the distribution of the observable random variables is known in each of the categories, except possibly for nuisance parameters. As a second possibility one may assume instead of known distributions that measurements of a number of individuals of known category (for example of known genotype with respect to some simple gene) are available. Thus in the case $k = 2$ we may have a number of Y's and Z's and want to classify an X as belonging to the same category as either the Y's or the Z's. For this problem it has been customary in the literature to consider the classification only of a single X. However, it seems clear that if several X's are to be classified the procedure can be much improved.

This becomes particularly clear if we go one step further and assume the distributions to be unknown but no Y's and Z's available for guidance. Suppose, for example, that $X_1, \cdots, X_n$ are each known to come from one of two normal distributions with $E(X_i) = \theta_i$ equal to either a or b and $\sigma^2_{X_i} = 1$, but that a and b are unknown. Suppose further that the θ's are independent random variables $P(\theta_i = a) = p$. It is then clear that for large n one will be able to obtain good estimates of a, b, p and hence carry out a reasonable classification of the X's. The problem is closely related to that of testing for outlying observations, which however is usually treated under somewhat different assumptions (see for example Grubbs [4] and Dixon [5]).

In the present paper we shall assume that all items with whose classification we are concerned are to be classified simultaneously. This is of course not always the case. Frequently the classification has to be carried out serially. It seems likely that in many cases the optimum serial classification procedure consists in classifying π_n on the basis of observations on $\pi_1, \cdots, \pi_n$ as if the problem were the simultaneous classification of $\pi_1, \cdots, \pi_n$. Hence the work done here should be applicable at least in part also to this problem.

In the present paper we shall consider the classification of individuals. For some simple problems the minimax procedures are obtained. Since they become asymptotically inadmissible as the number of individuals gets large other procedures are given that in the limit are minimax and admissible.

In a second paper we hope to treat analogous problems for the classification of populations. However here, as has already been pointed out, we shall keep the number of populations fixed and consider the case of large samples from each of the populations. This problem seems to be very much harder than the one treated here, and it is not too clear what general results to expect.

The present paper and the projected paper on the classification of populations are both related to a third paper on the theory of selection. In the first two papers it is assumed throughout that the categories are defined in absolute terms. The third paper constitutes an attempt at a problem in which this is not the case. It is concerned with classifying each of s populations as good or bad, where a population is defined as good if its quality is within given limits of that of the best of the populations. Although the theory of the minimax procedures is, as usual, easy (it involves an extension of the fundamental lemma of Neyman and Pearson to the case of vectorvalued critical functions) the application to particular cases presents difficulties which the author has not been able to overcome.

I should like to acknowledge my indebtedness to Dr. Howard Levene. In a seminar talk he presented his binomial example of the phenomenon discovered by Robbins and contrasted this with the minimax procedure. While much of the present paper was already written at the time, Dr. Levene's remarks suggested certain extensions of the work in progress.

A complete copy of Robbins' paper [1] was not available to me until after the present paper had been completed. In [1] Robbins indicates a method of approach which would seem to yield results, analogous to the ones obtained here, for a much more general class of problems.

2. Effect of a priori probabilities.

One of the main characteristics of problems concerning the classification of individuals is the assumption of a priori probabilities for the various categories. At first thought it might appear that the problem of the simultaneous classification of several individuals, as compared with the classification of one individual at a time, is much complicated by the fact that the a priori probabilities introduce dependence. Whether or not this is so depends, as has been pointed out by Mood [7] in a somewhat different context, on the procedure by which the individuals being classified have been obtained from the population. If the method is that of random sampling there is no dependence. This was shown by Mood for variables taking on only the values 1 and 0 and clearly holds in general. For let $X_1, \cdots, X_n$ be independently and identically distributed, and let $(i_1, \cdots, i_m)$ be m integer chosen at random from the set $(1, \cdots, n)$. Then the set of variables $(X_{i_1}, \cdots, X_{i_m})$ is clearly independent of the set of remaining X's.

If on the other hand the group being classified has been obtained by some other method, one will in general expect

dependence. It follows that for the problems under consideration it is important to know how the group that is being tested was obtained.

When we are dealing with a random sample from a population and if the proportion of the various categories in the population are known, it is clear from the above remark that the best method of simultaneous classification of the individuals in the sample consists in simply classifying each individual separately as best as possible without any regard to the remainder of the sample.

If, however, the proportions in the various categories are not known it has always been recognized that one should estimate these proportions from the sample. (See for example [6].) This is closely connected with the very interesting results obtained by Robbins [1]. He also pointed out that in the problem considered by him the minimax solution makes no use of the information that the sample contains concerning these proportions and that consequently the minimax solution is very inefficient for large samples. Another interesting example of the same nature was studied by Levene. The results of Robbins and Levene are considerably more startling than the ones we shall find here, since in their examples there is no assumption of any a priori probabilities. On the other hand their results are more difficult to interpret since there the parameter space changes with the sample so that there is no clear-cut fixed frame of reference.

3. An example.

Suppose we are interested in classifying n plants according to their genetic composition with respect to a single gene (a,A). We assume that the joint distribution of certain measurements is known for each of the possible genotypes, and is the same for dominants and hybrids. The plants come from the cross of a hybrid with a plant that was either recessive or hybrid. Hence it is known that the plants under consideration constitute a sample from a binomial distribution where the probability p of any one plant being recessive is either 1/4 or 1/2.

As throughout the paper we assume that the loss resulting from wrong classifications is measured by the number of these incorrect decisions, so that we want to minimize the expected number of misclassifications. If we adopt the minimax point of view it is easily seen that we shall act as if p were known to have that one of the values 1/4, 1/2, say p_0 which has the greater Bayes risk. (The Bayes risk corresponding to a value p_0 of p is the minimum expected number of misclassifications that can be achieved when p is known to be equal to p_0.) Each plant is then classified without regard to the measurements on the other plants in such a way as to maximize the probability of its correct classification.

Clearly, if n is at all large this procedure is very unreasonable. For we can then determine with near certainty whether $p = 1/4$ or $p = 1/2$. In one case we shall proceed as before, while in the other we shall modify our pro-

cedure. As a result of this modification there will of course be a slight increase in risk when $p = p_0$. This stems from the fact that there is a small but positive probability of having decided on the wrong value of p. However, this increase is balanced by a very substantial decrease in risk when p has the other value.

Let us now consider the problem quantitatively. We are concerned with random variables (presumably vectorvalued): $X_1, X_2, \cdots, X_n$. The variable X_i has probability density $p_{\theta_i}(x)$. The θ's are independent random variables, each capable of taking on the value 1 or 0. The probability $p = P(\theta_i = 1)$ is independent of i, and it is known that either $p = 1/2$ or $p = 1/4$. The problem is to classify each X_i into category $\mathcal{C}_1$ or $\mathcal{C}_0$ as θ_i is 1 or 0. If p were known, we would classify X_i into $\mathcal{C}_1$ if

$$\frac{p_1(x_i)}{p_0(x_i)} > \frac{q}{p}$$

and into $\mathcal{C}_0$ if the opposite inequality holds. The expected number of misclassifications in this case is $\alpha_p \cdot n$ where

$$\alpha_p = q\, P_0\left(\frac{p_1(X)}{p_0(X)} > \frac{q}{p}\right) + p\, P_1\left(\frac{p_1(X)}{p_0(X)} < \frac{q}{p}\right)$$

The minimax procedure clearly is the one appropriate to that

value p_0 of p (1/4 or 1/2) for which α_p is larger. To be specific, let us assume that $p_0 = 1/2$.

Let Y be the number of variables X_i that are classified into ζ_1 by the minimax procedure. Then Y is the number of successes in n independent trials with constant probability

$$p(1-\alpha) + (1-p)\,\alpha = \alpha + p(1-2\alpha)$$

of success. Hence $\dfrac{\frac{Y}{n} - \alpha}{1 - 2\alpha}$ is a consistent estimate of p.

Let us replace the minimax procedure by the following:

If $\dfrac{\frac{Y}{n} - \alpha}{1 - 2\alpha} > \dfrac{3}{8}$ use the minimax procedure

If $\dfrac{\frac{Y}{n} - \alpha}{1 - 2\alpha} < \dfrac{3}{8}$ use the procedure appropriate for p = 1/4.

To compute the risk of this new procedure suppose first that p = 1/2. Then the expected number of misclassifications is

$$(1) \quad P\left(\frac{\frac{Y}{n}-\alpha}{1-2\alpha} > \frac{3}{8} \,\middle|\, p = \frac{1}{2}\right) \frac{n}{2}\left[P_0\left(\frac{p_1(X)}{p_0(X)} > k \,\middle|\, \frac{\frac{Y}{n}-\alpha}{1-2\alpha} > \frac{3}{8}\right)\right.$$

$$+ P_1\left(\frac{p_1(X)}{p_0(X)} < k \,\middle|\, \frac{\frac{Y}{n}-\alpha}{1-2\alpha} > \frac{3}{8}\right)\Bigg] + P\left(\frac{\frac{Y}{n}-\alpha}{1-2\alpha} < \frac{3}{8} \,\middle|\, p = \frac{1}{2}\right) n\,\beta_n$$

where $\beta_n \leq 1$.

It is easily seen that

$$n\,P\left(\frac{\frac{Y}{n}-\alpha}{1-2\alpha} < \frac{3}{8} \,\middle|\, p = \frac{1}{2}\right) \longrightarrow 0.$$

Hence the second term of (1) tends to zero. In the first term, the first factor tends to 1, while the last factor tends to the sum of the unconditional probabilities. Thus the ratio of the risk to the minimax risk tends to 1 as $n \longrightarrow \infty$.

An exactly analogous argument shows that when $p = 1/4$ the ratio of the new risk to the risk of the Bayes procedure corresponding to $p = 1/4$ also tends to 1.

We have used here for simplicity the frequency Y/n to decide between $p = 1/2$ and $p = 1/4$. However more sensitive methods are available, and one should expect these to yield better results also for the classification problem. Thus one might decide for $p = 1/2$ if

$$\sum_{i=1}^{n} \log \frac{p_0(X_i) + p_1(X_i)}{p_0(X_i) + 3p_1(X_i)} > k$$

where k is some suitable constant.

It should also be pointed out that although the procedure discussed here has good asymptotic properties, it is not admissible. In fact it is easy to obtain the totality of admissible procedures since by Wald's theory [8] this coincides in our case with the totality of Bayes solutions. But there are only two parameter points: $p = 1/2$ and $p = 1/4$. Hence a Bayes formulation assumes probabilities $\rho = P(p=\frac{1}{2})$, $1-\rho = P(p=\frac{1}{4})$, and the class of all Bayes solutions is a one-parameter family, one solution for each value of ρ.

These Bayes solutions are easy to obtain as follows. Any classification procedure of n items into two categories is a vectorvalued function $\phi(x) = (\phi_1(x), \cdots, \phi_n(x))$, $0 \leq \phi_i(x) \leq 1$. If x is observed, the i-th item is classified into ζ_1 with probability $\phi_i(x)$, into ζ_0 with probability $1 - \phi_i(x)$. Instead of minimizing the Bayes risk of a procedure ϕ corresponding to some given value ρ, we shall maximize the expected number of correct decisions, which is given by

$$\rho \sum_{k=0}^{n} \sum_{(i_1, \cdots, i_k)} \binom{n}{k} \frac{1}{2^n} E\left[\Phi_I\right]$$

$$+ (1-\rho) \sum_{k=0}^{n} \sum_{(i_1, \cdots, i_k)} \binom{n}{k} \frac{1}{4}^{k} \left(\frac{3}{4}\right)^{n-k} E\left[\Phi_I\right]$$

where $E\left[\Phi_I\right]$ denotes

$$E\left[\Phi\right] = E\Big[\phi_{i_1}(X) + \cdots + \phi_{i_k}(X) + 1-\phi_{j_1}(X) + \cdots$$

$$+ 1-\phi_{j_{n-k}} \Big| \theta_{i_1} = \cdots = \theta_{i_k} = 1,\ \theta_{j_1} = \cdots = \theta_{j_{n-k}} = 0\Big]$$

with the summation $(i_1, \cdots, i_k)$ extending over all combinations of k integers out of $(1, \cdots, n)$ and where $j_1, \cdots, j_{n-k}$ denotes the remaining integers.

Thus we have to maximize an expression of the form

$$\sum_i \int \left[\sum a_{ij}\phi_j(x)\right] p_i(x)\, d\mu(x)$$

which is achieved by setting

$$\phi_j(x) = 1 \qquad \text{whenever} \quad \sum_i a_{ij}\, p_i(x) > 0$$

$$\phi_j(x) = 0 \qquad \text{whenever} \quad \sum a_{ij}\, p_i(x) < 0.$$

Unfortunately, although it is easy to write this down explicitly, the resulting procedure does not seem very manageable.

4. Asymptotic theory.

The inefficiency for large samples of the minimax solution, that we found here is not an isolated phenomenon.

There are many cases of a sequence of problems Π_n for each of which a minimax solution exists that is unique and hence admissible. However for large n there exist other procedures which at the expense of a slight increase of the risk functions for some parameter values reduces the risk very substantially for other values of the parameter.

These considerations lead to the following definition. Let the distribution of the observable random variables depend on a parameter θ, and denote the risk function of a decision procedure δ by $R_\delta(\theta)$. Then we shall say that a sequence of decision procedures is asymptotically non-admissible if there exists a sequence of procedures δ_n^* such that

$$(1) \qquad \overline{\lim_{n \to \infty}} \frac{R_{\delta^*_n}(\theta)}{R_{\delta_n}(\theta)} \leq 1 \quad \text{for all} \quad \theta$$

with strict inequality holding for some θ. (The results of section 2 show that in the example considered there the minimax procedure is asymptotically non-admissible.) In analogy with the above definition one can define the notion of asymptotic admissibility and an asymptotic minimax procedure; this latter notion was introduced by Wald [9]. It also seems useful to define the following concept:

A sequence of procedures $\delta_n^{(o)}$ is said to be consistent if for each θ

$$\lim_{n\to\infty}\left[R_{\delta_n^0}(\theta) - \inf_{\delta_n} R_{\delta_n}(\theta)\right] = 0$$

We can then state the following obvious result. If for each θ

$$\lim_{n\to\infty}\left(\inf_{\delta_n} R_{\delta_n}(\theta)\right) > 0$$

and if δ_n^0 is consistent, then δ_n^0 is asymptotically admissible and minimax.

We shall now consider the asymptotic theory of the simultaneous classification of a large number of individuals. Let X_i be independently distributed with density $p_{\theta_i}(x)$ where the θ_i are independent random variables taking on the values 1,0 with probabilities p,q respectively. It is desired to classify each X_i according to its θ_i in such a way as to minimize the expected number of misclassifications.

For simplicity assume that $\frac{p_1(X)}{p_0(X)}$ has a continuous distribution both when θ_i is 1 and 0. The minimax procedure classifies X_i into $\mathcal{C}_1$ if and only if

$\frac{p_1(X_i)}{p_0(X_i)} > k$, where k is determined by the condition

$$P_0\left(\frac{p_1(X)}{p_0(X)} > k\right) = P_1\left(\frac{p_1(X)}{p_0(X)} < k\right) \quad (= \alpha, \text{ say}) \quad \alpha < \frac{1}{2}.$$

Theorem.

Let Y be the number of X's that are classified ζ_1 by the minimax procedure. Then as $n \to \infty$ the following sequence of procedures is consistent and hence asymptotically admissible and minimax.

Classify X_i into ζ_1 if and only if

$$\frac{p_1(X_i)}{p_0(X_i)} > \frac{(1-\alpha) - \frac{Y}{n}}{\frac{Y}{n} - \alpha}$$

Proof. We note first that $\dfrac{(1-\alpha) - \frac{Y}{n}}{\frac{Y}{n} - \alpha}$ is a consistent estimate of q/p. For Y/n is the frequency of success in n independent trials with constant probability

$$q\, P_0\left(\frac{p_1(X)}{p_0(X)} > k\right) + p\, P_1\left(\frac{p_1(X)}{p_0(X)} > k\right)$$

$$= q\,\alpha + p(1-\alpha) = \alpha + p(1-2\alpha)$$

of success. Therefore $\dfrac{\frac{Y}{n} - \alpha}{1 - 2\alpha}$ tends in probability to p.

But

$$\frac{(1-\alpha) - \frac{Y}{n}}{\frac{Y}{n} - \alpha} = \frac{1 - 2\alpha}{\frac{Y}{n} - \alpha} - 1$$

and hence tends in probability to

$$\frac{1}{p} - 1 = \frac{q}{p}$$

Now the expected number of misclassifications when p is true and we use the proposed procedure is

$$n\left[q\ P_0\left(\frac{p_1(X_1)}{p_0(X_1)} > \frac{(1-\alpha)-\frac{Y}{n}}{\frac{Y}{n}-\alpha}\ \middle|\ p\right) + p\ P_1\left(\frac{p_1(X_1)}{p_0(X_1)} < \frac{(1-\alpha)-\frac{Y}{n}}{\frac{Y}{n}-\alpha}\ \middle|\ p\right)\right]$$

where P_0, P_1 indicate the distribution of X_1 while it is assumed that p is the probability of θ_i being 1.

If p were known we could use the above procedure with the quantity $\frac{(1-\alpha)-\frac{Y}{n}}{\frac{Y}{n}-\alpha}$ replaced by q/p. Hence in order to prove our result we need only show that as $n \to \infty$

$$(2)\quad P_i\left(\frac{p_1(X_1)}{p_0(X_1)} > \frac{(1-\alpha)-\frac{Y}{n}}{\frac{Y}{n}-\alpha}\ \middle|\ p\right) \to P_i\left(\frac{p_1(X_1)}{p_0(X_1)} > \frac{q}{p}\right) \text{ for } i = 1, 0.$$

Now let Y' be the number of X's among $X_2, \cdots, X_n$ that satisfy $\frac{p_1(X)}{p_0(X)} > k$. Then $|Y'-Y| \leq 1$ and it is clear that we can replace the left hand side of (2) with

$$P_i\left(\frac{p_1(X_1)}{p_0(X_1)} > \frac{(1-\alpha)-\frac{Y'}{n}}{\frac{Y'}{n}-\alpha}\right).$$

But X_1 and Y' are independent, and the result follows from the following fact:

If X, U_n are independent, $U_n \to a$ in probability, and a is a continuity point of X, then

$$P(X > U_n) \to P(X > a).$$

It should be pointed out that the theorem would presumably remain valid if we replaced our estimate of q/p by any other consistent estimate. The next stage would be to consider the speed with which the limiting risk is approached, as one uses different estimates of q/p.

We have stated the theorem for the case of only two categories. The extension to the case of s categories is easy and we shall only sketch it briefly. We now assume that each θ_i can take on s values, say $1, 2, \cdots, s$, and that each X_i is to be classified into one of the classes $C_1, \cdots, C_s$ according to the value of θ_i. If π_j is the a priori probability of any one θ taking on the value j, the associated Bayes solution classifies an X into C_i if

$$\pi_i \, p_i(x) = \max_j \{\pi_j \, p_j(x)\}. \tag{3}$$

Let R_i be the set of points x for which (3) holds, and let

$$\alpha_{ij} = P_j(X \in R_i).$$

If Y_i is the number of X's classified as C_i under the

Bayes procedure associated with some particular set of a priori probabilities $(\pi_1, \cdots, \pi_s)$, then Y_i is the number of outcomes i in n multinomial trials in which the outcome i has constant probability

$$\sum_{j=1}^{s} \pi_j \alpha_{ij} .$$

Assume that $\pi_i > 0$ for $i = 1, \cdots, s$ and let $(\hat{\pi}_1, \cdots, \hat{\pi}_s)$ be the solutions of the system of equations

$$\sum_{j=1}^{k} \hat{\pi}_j \alpha_{ij} = \frac{Y_i}{n} .$$

Then the procedure that classifies X_r into C_i if

$$\hat{\pi}_i p_i(X_r) = \max_j \hat{\pi}_j p_j(X_r)$$

is consistent.

5. Stratification.

It frequently happens that there is more information available than was assumed in the last section. Suppose namely that the population is stratified (for example by sex, age, previous training, etc.) We still have to classify each individual into one of a number of classes, however the proportions of individuals in the various categories presumably differ among the various strata.

Let us consider the simplest case of two categories and two strata. The individuals of the two strata will be de-

noted by Y and Z respectively. Each Y_i has a probability density $p_{\eta_i}(y)$, $\eta_i = 0$ or 1; each Z_i has a density $p_{\xi_i}(Z)$, $\xi_i = 0$ or 1. Let us put

$$P(\eta_i = 1) = p$$

$$P(\xi_i = 1) = p'$$

and let π be the proportion of Y's in the total population. It is clear that if $p \neq p'$ the procedure discussed in the last section, which takes no account of the stratification, will lose its asymptotic properties. For let us assume for a moment that p, p', π are known. Then the (unique) optimum procedure will differ in its treatment of the Y's and Z's. On the other hand the procedure that minimizes the risk under the assumption that all individuals are drawn at random from a population with a proportion $\pi p + (1-\pi) p'$ of individuals belonging to class C_1, will classify all of the individuals according to the same rule, and hence has a higher risk. Since the asymptotic risk, when the various parameters are estimated, has been shown to approach the Bayes risk when they are known, the result follows.

It is also clear now that in the present case the following procedure will be consistent. We estimate separately p and p', say by $\hat{p}$ and $\hat{p}'$ and classify a Y into C_1 if

$\frac{p_1(Y)}{p_0(Y)} > \hat{p}$, and a Z into C_1 if $\frac{p_1(Z)}{p_0(Z)} > \hat{p}'$. As a matter of fact this procedure retains consistency even when $p = p'$, however this is an indication of the weakness of the definition rather than the quality of the procedure.

Taking account of stratification not only improves the risk function but it also avoids the necessity of making assumptions about how the sample is divided between the strata. In the first treatment we assumed random sampling from the total population. However if the stratification is such that the strata differ markedly in the proportions of the various categories this assumption is likely to be fallacious.

While thus from a statistical point of view there are considerable advantages in not using the minimax procedure, this, at least in certain problems, also entails serious disadvantages of an ethical nature. While the issue is brought out particularly clearly in connection with stratification, it is actually present in the whole discussion. If each individual can be either 0 or 1 and if some significance attaches to the classification, one feels strongly that each person should be classified on the basis of his own performance without regard to that of the other individuals being classified.

At first one may feel that the fault lies with the loss function. We have stated it as our task to minimize the total average number of misclassifications. However exactly the same phenomenon occurs if we are interested in classifying only

individual i. If we want to minimize the probability of misclassifications we will estimate the proportion of 0's in the population, and proceed as we did before.

The moral conflict arises with the assumption of random sampling from the population. The individual does not consider himself drawn at random from a population. For him θ is not a random variable but a parameter. Thus, if we want to meet this objection we have to forego the advantage of the assumption of random sampling and treat the θ's as parameters.

It might seem that even then the difficulty remains because of the possibilities brought out by Robbins. This is however not so. The phenomenon described by Robbins occurs if we express the risk in terms of the frequency of θ's in the group that is being classified. But this is inappropriate if we are concerned with the classification of the single individual. Then θ_i is 0 or 1, and the risk must be expressed in terms of these two possibilities and not an extraneous frame of reference.

6. On a general class of problems.

The problems discussed here have certain features in common with a large class of statistical problems. As we shall indicate, it seems likely that the results we obtained in the special cases apply more generally.

In the examples we considered the distribution of the observable random variables was, as is usually the case, only partially known. However -- and here they differ from

the classical problems of statistics -- even if the distribution were known this would still not imply knowledge of the correct decision since this depends on the values of some unobservable random variables. The same situation occurs, for example, in all prediction problems.

Suppose in general that we are concerned with a situation in which a decision is to be made on the basis of observable random variables $X_1, \cdots, X_n$ whose joint distribution, for all n, depends on a certain parameter θ. It is assumed that as $n \to \infty$ one can estimate θ consistently. The correct decision depends not directly on θ but on certain unobservable random variables the distribution of which also involves θ.

In such cases it seems to be true rather generally that the minimax procedure is asymptotically inadmissible.

For suppose that θ were known and let θ_0 be that value of θ to which corresponds the biggest (Bayes) risk. In some cases the minimax solution is, for all n, the Bayes solution corresponding to this worst value of the parameter. (This is the case in the example considered in section 2 and in the prediction of the outcome of a single binomial trial (see [10]).) In other cases the least favorable distribution (if one exists) is not concentrated at this one value since it takes into account both the difficulty of determining the correct value of θ and the difficulty of determining the correct value of the non-observable variables when θ is known. However, as $n \to \infty$, the difficulty of determining the correct value of θ gradually disappears, and hence the least favorable distribution

does tend to concentrate around the "worst" value of θ. The asymptotic inadmissibility of the Bayes solution corresponding to the least favorable distribution (i.e., of the minimax procedure) now follows as before from the fact that we can determine the time value of θ quite accurately and hence do not have to take the pessimistic attitude of the minimax solution.

At the same time we see that the Bayes procedure corresponding to the estimated value $\hat{\theta}$ of θ is consistent and hence asymptotically admissible and minimax. For the contribution to the risk resulting from having the wrong value of θ tends to zero, and hence the total risk tends to the risk one would be left with even if θ were known.

In some problems of the kind being considered one can avoid the difficulties of the minimax procedure by adopting the notion of L.J. Savage, of minimizing the maximum regret. It is easy to see, for example, in prediction problems with squared error as loss function that the prediction of a random variable that minimizes the maximum regret is the same as the minimax estimate of $E(Y)$.

It should finally be pointed out that the asymptotic inadmissibility of minimax procedures may also occur in classical problems where the difficulties discussed in the present section do not arise. An example in question is the estimation of a binomial probability [10].

References

[1] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems", Second Berkeley Symposium.

[2] H. Levene, unpublished.

[3] William C. Boyd, "Genetics and the races of man", Little, Brown and Company, Boston, 1950.

[4] F.E. Grubbs, "Sample criteria for testing outlying observations", Annals of Natl. Stat., Vol. 21 (1950), p. 27.

[5] W. Dixon, "Analysis of extreme values", Annals of Math. Stat., Vol. 21 (1950), p. 488.

[6] P.G. Hoel and R.P. Peterson, "A solution to the problem of optimum classification", Annals of Math. Stat., Vol. 20 (1949), p. 433.

[7] A.M. Mood, "On the dependence of sampling inspection plans upon population distributions", Annals of Math. Stat., Vol. 14 (1943), p. 415.

[8] A. Wald, "Statistical decision functions", Wiley and Sons, New York, 1950.

[9] A. Wald, "Asymptotic minimax solutions of sequential point estimation problems", Second Berkeley Symposium.

[10] J.L. Hodges, Jr. and E.L. Lehmann, "Some problems in minimax point estimation", Annals of Math. Stat., Vol. 21 (1950), p. 182.

USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 6.

Discriminatory Analysis: VI. On the Simultaneous Classification of Several Individuals.

E.L. Lehmann, University of California, Berkeley.

27 pp & i. 0 illus. 27 cm. UNCLASSIFIED

The classification of individuals is considered, and for some simple problems minimax procedures are obtained. Other procedures are given which are minimax and adminissible in the limit as the number of individuals becomes larger.

1. Biometrics 2. Statistical Analysis I. Lehmann, E.L.

USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 6.

Discriminatory Analysis: VI. On the Simultaneous Classification of Several Individuals.

E.L. Lehmann, University of California, Berkeley.

27 pp & i. 0 illus. 27 cm. UNCLASSIFIED

The classification of individuals is considered, and for some simple problems minimax procedures are obtained. Other procedures are given which are minimax and admissible in the limit as the number of individuals becomes larger.

1. Biometrics 2. Statistical Analysis I. Lehmann, E.L.